Chloroinformatics

Ali Amiryousefi

Organismal Evolutionary Biology Research Program
Faculty of Biological and Environmental Sciences
University of Helsinki
Finland

&

Botany Unit
Finnish Museum of Natural History
University of Helsinki
Finland

ACADEMIC DISSERTATION

To be presented for public examination with the permission of the Faculty of Biological and Environmental Sciences of the University of Helsinki in the Nylander hall, Kaisaniemi Botanical Museum (Unioninkatu 44) on 19th of March at 12:00.

Helsinki 2019

Supervised by: Dr. Péter Poczai
Finnish Museum of Natural History
University of Helsinki, Finland

Prof. Jaakko Hyvönen
Finnish Museum of Natural History &
Organismal Evolutionary Research Program
University of Helsinki, Finland

Reviewed by: Prof. Françoise Budar
Institut Jean-Pierre Bourgin
French National Institute for Agricultural Research, France

Dr. Endre Barta
Department of Genomics
Agricultural Biotechnology Centre, Hungary

Examined by: Prof. Ildikó Karsai
Department of Molecular Breeding
Hungarian Academy of Sciences, Hungary

Custos: Prof. Jouko Rikkinen
Organismal Evolutionary Research Program
University of Helsinki, Finland

Advisory Committee: Prof. Eva-Mari Aro
Department of Biochemistry
University of Turku, Finland

Prof. Teemu Teeri
Department of Agricultural Sciences
University of Helsinki, Finland

ISBN 978-951-51-5102-5 (paperback)
ISBN 978-951-51-5103-2 (PDF)
http://ethesis.helsinki.fi
Helsinki, 2019

Contents

This thesis is based on the following articles, which are referred to in the text by their Roman numerals.

I **Amiryousefi, A.,** Hyvönen, J., Poczai, P. 2017. The chloroplast genome sequence of bittersweet (*Solanum dulcamara*): genome structure evolution and phylogenetic relationships in Solanaceae. *PLoS One* 13(4): e0196069.

II Sablok, G., **Amiryousefi, A.,** He, X., Hyvönen, J. Poczai, P. 2019. Sequencing the plastid genome of giant ragweed (*Ambrosia trifida*, Asteraceae) from the herbarium specimen. *Frontiers in Plant Sciences* 10(218).

III **Amiryousefi, A.,** Hyvönen, J., Poczai, P. 2018. IRscope: An online program to visualize the junction sites of the chloroplast genomes. *Bioinformatics* 34(17): 3030-3031.

IV **Amiryousefi, A.,** Hyvönen, J., Poczai, P. 2018. *i*MEC: Online marker efficiency calculator. *Applications in Plant Sciences* 24(6): e01159.

Table of contributions

	I	**II**	**III**	**IV**
Original idea	PP	PP	**AA**	**AA**, PP
Analysis	**AA**, JH, PP	**AA**, GS, PP, JH, XH	**AA**, PP	**AA**
Software	-	-	**AA**	**AA**
Manuscript Preparation	**AA**, JH, PP	**AA**, GS, PP, JH, XH	**AA**, JH, PP	**AA**, JH, PP

AA **= Ali Amiryousefi**
JH = Jaakko Hyvönen
PP = Péter Poczai
GS = Gaurav Sablok
XH = Xiaolan He

... to my roots.

Abstract

Chloroplasts are cytoplasmic organelles chiefly responsible for the photosynthesis. Their genes have been used extensively during the past decades in phylogenetic analyses of various photosynthetic eukaryotes, particularly plants. The genomic content of this organelle and its very architecture can be used for a deeper insight in evolution and towards robust phylogenetic hypotheses. Ever since this importance was recognized concurrently with the advancements of methods in both providing a basic genetic material through sequencing and advanced methods to analyze the data, we have witnessed the introduction of a couple of thousands plastid genomes up to this date. This process, by no means is in its decline or even stationary state, as this pace is projected to be accelerated in the coming years, with the inevitable advances in our technologies and our need to understand the nature as accurately as possible. The aim of this study as represented in the sequel chapters is twofold; 1) to introduce the complete chloroplast genomes of two species from the euasterid clade and provide their phylogenetic analyses; *Solanum dulcamara* L. as a native Old World diploid member of the nightshade family, and *Ambrosia trifida* L. as a recognized invasive plant originated and evolved from North America. 2) To provide two analytical tools for more advanced treatment of the genetic information of plastids in bioinformatics. By comparative analyses for bittersweet and giant ragweed, the result show that synteny and the genomic content of both belonging to the families Solanaceae and Asteraceae, respectively, have a conserved structure. We also noted that many submitted annotations in the nightshade family are far from acceptable quality, and further on, we improved them with reannotation of the existing sequences. On the other hand, a novel tool (IRscope) to detect and plot the Inverted Repeat (IR) regions of the chloroplast genome was introduced. IRscope, with the help of iterative search algorithm, allows the depiction of genes in the vicinity of the Junction Sites (JS), of up to ten different chloroplast genomes of embryophytes (land plants). Moreover, we constructed an online calculative suite (iMEC) to return the result of the seven different molecular markers against the provided input file. This tool is useful particularly in studies aimed to assess the efficiency of different marker systems linked to plastid genome variation.

Summary

Ali Amiryousefi
ali.amiryousefi@helsinki.fi

1. Introduction

"We cannot discover new lands, without consenting
to lose sight from the shore for a long time"

- André Gide

The diversity of life on our planet with almost nine million extant species, probably would have not been possible without oxygen (Mora *et al.,* 2011). The oxygenic photosynthesis in cyanobacteria and the later developed aerobic respiration (ca. 2.3 billion years ago) as a complex multicellular machinery are the phenomenal incidents tied with this historical landmark (Soo *et al.*, 2017). While there is evidence that anoxygenic photosynthesis emerged not too long after the origin of life on Earth (Blankship, 2010), photosynthesis as we know it today was an unprecedented event that led to the accumulation of oxygen and diversification of life. Although much challenged with a set of different hypotheses about the origin and exact placement of the engulfing event on the tree of life (Degan *et al.,* 2013; Ponce-Toledo *et al.,* 2017; Sanchez-Baracaldo *et al.,* 2017; Rodriguez-Ezpeletaet *et al.,* 2005; Blank & Sanchez-Baracaldo, 2010; Criscuolo *et al.*, 2011; Uyeda *et al.*, 2016), recent evidence indicate that it took some million years for organellar photosynthesis to evolve from the so-called primary endosymbionts about one billion years ago (Zhang *et al.,* 2018 & Betts *et al.,* 2018). During this gradational process loss or translocation of the majority of the genes of the endosymbionts to the host nuclear genome took place. But this transaction was not only in one direction as the protein products of many of these genes were later to be reimported into the plastids by the sophisticated apparatus called the TIC-TOC complex (translocon complex of the inner and outer chloroplast membranes; Nakayama and Archibald, 2012).

Plastids, ubiquitously present in the various plants and algae (Whatley 1978; Keeling 2010), are the light-harvesting organelles of photosynthetic eukaryotes and are believed to be derived from once free-living cyanobacteria by a specific process of endosymbiosis (Lane and Archibald, 2008). This process has been responsible for some of the most significant events in evolution of eukaryotic cells. A great body of biochemical and molecular data hint that a prokaryotic relative of extant cyanobacteria was engulfed and further retained in time by a heterotrophic eukaryote (Deschamps *et al.*, 2008). This progressively transformed into a photosynthetic organelle of plants and various other eukaryotes. Integration of a prokaryotic

endosymbiont into the cellular machinery of a eukaryote, is a complicated process including substantial modifications to the genetic makeup of both participating cells (Bhattacharya *et al.*, 2007). Virtually all known organelles originated through endosymbiosis entail only a fraction of the genes as compared to their prokaryotic closest relatives. This indicates that the majority of genes that were once essential to the free-living prokaryote (but obsolete in the intracellular context), are either transferred to the host nucleus, or lost by the genomic degradation. The length of plastid genomes is considerably less than genomes of most free-living cyanobacteria (Stoebe and Kowallik, 1999). As the genetic capabilities of the prokaryotic endosymbiont has diminished during the transition from a free-living cell to a fully engulfed organelle, the host cell became a repertoire of genetic information through endosymbiotic gene transfer (EGT; Martin *et al.*, 2002). Many of the genes transferred to the nucleus of the host cell acquire targeting signal capability, which enable their products to be transported back to the plastid to perform vital functions (Bhattacharya *et al.*, 2007). However, transferred genes can also obtain novel functions in the eukaryotic cell, and sometimes even replace the eukaryotic versions of the proteins they encode. It is a general consensus that in Eukaryota, plastids evolved from cyanobacteria multiple times during the history of life (Larkum *et al.*, 2007; Stiller *et al.*, 2003; & Archibald 2009). And this process was further complicated with multiple different losses of parts to the nucleus and engulfing again in future (Archibald, 2009). These intricate and multiple events in history of the plastids renders the precise pinpointing of these timings in evolution a challenging task (Nozaki and Iseki, 2007; Yoon *et al.*, 2004, & Stiller 2007). Regarding he placement of the plastids in the extant diversity of Cynobacteria, one hypothesis is postulating the ancient divergence (Ponce-Toledo *et al*., 2017; Sanchez-Bracaldo *et al*., 2017; Criscuolo *et al.*, 2011; & Li *et al.*, 2013), while the other postulates the relatively recent origin (Ochoa de Alda., 2014; Deschamp *et al.*, 2008; & Falcon *et al*., 2010) but after all, the factual placement of plastid on the cynobacterial lineages is still unresolved.

The evolutionary forces albeit were not restricted in action at the molecular level. The environment has had played an important role in diversifying or eliminating various organismal lineages until to date. After all, we have now projected 300,000 plant species (although argued to be higher by 10-20%, Joppa *et al.*, 2010) alive that have survived the estimated 99% extinction of all life through the history (Mora *et al.*, 2011, Novacek and Wheeler, 1992). While morphologically more than 85% of these plant species have been described, our genomic knowledge about them is still sparse. After the first whole genome reconstruction in 1995, we only have ~240 full genome sequences of the plants (Archaeplastida; Chen *et al.*, 2018). This value is nominal (comprising less than 0.1%) in comparison to the total number of plant species. The complete plastid genomes of Eukaryota sequenced today on the other hand, is 2,946 as of 08.01.2019, NCBI Organellar Genome Database. The first plastid genome sequences are from 1986 when *Nicotiana tabacum* L. and *Marchantia polymorpha* L. were sequenced (Ohyama *et al*, 1986; Shinozaki *et al*, 1986). This

precedence both in numbers and time of sequencing for plastid genomes can be due to the small size of these organelle sequences ranging from 75-200 kbp[1] across tree of life (Green, 2011); or the fact that the plastid genome is the most gene-rich of the three genomes in each cell hence they can be a reliable source for evolutionary studies; or because its copy number is the highest and its sequencing is the most cost effective (Wicke *et al,*. 2011); or, maybe simply because of the essential role that this organelle plays in orchestrating the most intricate and crucial function of plants, photosynthesis.

After all, ever since early sequencing methods, e.g., chain termination, technology has matured to high-throughput sequencing like sequencing by synthesis (illumina) and single molecule real time (SMRT) sequencing (PacBio), which has positively correlated with the exponential increase in the number and the quality of the sequences (Hug *et al*, 2016). This trend, projected to be maintained in the future, the challenge is to keep up with the invention of sophisticated bioinformatics methods honed in deducing the untapped potential of this new information. While diverse set of recommendations on how to scale our understanding with respect to this data has been put forward (e.g. Tonti-Filippini *et al.*, 2017), we seem not yet fully prepared to capture this large flow of genomic data (Wicke and Schneeweiss, 2015). Our respective shortcomings can be named as (Kiureghian & Ditlevsen, 2009); 1) the inability of our tools to perform as credibly as they have expected or promised to (epistemic bias), and 2) the nonexistence of crucial methods and theoretical knowledge to reliably infer the latent dependencies of intrinsic genomic complexity (aleatory bias).

As the first part of this study, we have presented two complete plastid genomes sequences belonging to Solanaceae and Asteraceae. The studies of comparative analysis encompass all the existing plastid sequences in these families. Analysis hence in totality enables us to address the underlying biodiversity of the euasterids in higher resolution. On the other hand, with invention of visualization and computational tools, the other part of this study is an effort in resolving the two methodological pitfalls mentioned above. This seemed to be a crucial next move in filling the gap between the incoming data and existing methods. This study is aimed to impact the enrichment of both resources and methods of plastid genomics. More precisely, with reference to the original contributions of this study on the material side, the next section of this dissertation discusses two different sequenced chloroplast genomes with the emphasis on those of *Solanum dulcamara* and *Ambrosia trifida* as presented in Chapter **I**

[1] This range is excluding the peculiarity of some species that can be categorized as outliers. The chloroplast genome of the *Haematococcus lacustris* L. (*MG677935.1*) belonging to Chlorophyta for example is 1,352,306 bp. Inflated with repeats, in general, Chlorophyta is exhibiting the longest plastid genome sequences ranging 124-521 kbp, while intracellular parasites and non-photosynthetic heterotrophic apicomplexans of Chromalveolata are on the other hand of the spectrum with 35-40 kbp long plastid genomes sequences (Green, 2011), with *Pilostyles aethiopica* L. (*NC_029235.1*) as the shortest plastid genome with only 11,348 bp and five functional genes (Bellot & Renner, 2016).

and **II**, respectively. This will be followed with the improvement and introduction of new methods to better assess the chloroplast genomes in the third section as discussed in Chapter **III** and **IV**. The fourth section discusses some of the shortcomings and concerns related to this study and delineates some possible future directions of research, and finally summarizes the main findings and implications.

2. Chloroplast evolution

"Each allied with me based on his belief,
But alas bereft from my secrets within"

- Rumi

We probably owe less to J. Priestley (1733-1804) the 1773 discovery of oxygen as nature does to evolution for its invention of the photosynthesis, ~2.3 billion years earlier (Soo *et al*., 2017). Ever since this invention in the cyanobacteria, the evolution bestowed this mechanism to Archaeplastida consisting of three groups of primary endosymbionts as green algae+plants, red algae, and glaucophytes (Blankenship 2009; Douglas 1998). Each group have slightly modified this process to best hone the function and survival of the species and despite the overall stability of photosynthesis, it is possible to find dissimilarities in the related underlying photosynthetic genotypes of different species, even at the genus level (Nevo *et al.*, 2012).
This section will mainly focus on two first chapters of this dissertation presenting the plastid DNA architecture and its evolutionary aspects in Solanaceae and Asteraceae. The presented chapters deliver a high-resolution analysis of the chloroplast genomes of the particular species as well as other analysis concerned with these genomes in relation to other relevant closely related species in the form of comparative genomics and phylogenetic analyses. The following section discusses the chloroplast as a main biological organelle regarded as a building block to our analyses. This will be followed with beckoning on some of the analyses performed with the genomes that has helped us in understanding the position of these species on the plant tree of life.

2.1 Plastid genome architecture

Chloroplasts are the prototypical members of a diverse family of organelles; the plastids. In plants, other plastid family members are the amyloplasts (found in seeds, roots and tubers) and chromoplasts (which accumulate carotenoid pigments and function as attractants in flowers and fruits; Lopez-Juez & Pyke, 2005). Proplastids are small, undifferentiated plastids that exist in meristems and reproductive tissues. The specific function of the proplastids is organelle transmission between generations and within all cells of the organism. The more specialized plastids in plants are derived from proplastids through differentiation and all are bounded by a double-membrane system that is called the envelope (Sakamoto *et al*,. 2008). Chloroplasts are mainly found in the cells of the mesophyll, the tissue in the interior of the leaf. A typical mesophyll cell contains about 35 chloroplasts each measuring ~16 μm^2. The chloroplast genome (cpDNA) is in stroma, a dense fluid enclosed by a double membrane separating the chloroplast from the cytosol. This cpDNA is undoubtedly a remnant of the evolutionary origin of the plastid as endosymbiotic cyanobacteria (Blankenship 2010).

The boldest differentiating aspect of the plastid genome as compared to that of cyanobacteria is its smaller size and hence, its reduced genetic contents. This reduction in size is in part[2], the result of the fact that most of the genes needed for photosynthesis has been transferred to the nucleus. The question then arises as, if there are advantages in this transfer of the genes to the nucleus, why not all of them? Two answers have been put forward to this end. The first is that certain plastid proteins are intrinsically difficult to be transported through the plastid envelope. Then it renders it difficult to translocate its underlying gene when consequent synthesis of proteins in the cytosol and the post-translational import into the organelle is concerned (Howe *et al.*, 2002). While this suggestion seems to be plausible, there are studies that show that under certain conditions, it is possible to artificially introduce the plastid genes to the nucleus and effectively re-import the resulting mature protein back in the organelle (Cheung *et al*., 1988; Kanevski & Maliga 1994). Given this re-importing can in principle be obtained, one might note minor reduction in fitness that this procedure brings about for the organism. The second suggestion for the retention of the plastid genes is that it allows the instantaneous regulation of the expression in response to the redox status of the organelle (Pfannschmidt *et al.*, 1999; Allen 1993). On the other hand, although lacking fully supportive indications, the result of specified tests tailored for assessing the colocation of the gene and the gene product for redox regulation of their expression, seem to be in favor of this hypothesis (Allen, 2015). Chloroplast genomes of the vascular plants normally possess 50 to 80 coding protein genes which reside on 120 - 200kb long genome sequence. These genes contain the information for many of the core proteins of the photosystems and the cytochrome b_6f complex (Green, 2011; Table 1). The number of proteins coded by a chloroplast genome range from 1000 to 5000, which are far smaller than what is needed for assembly of the photosynthetic complexes. This huge deficit indeed is supplemented with the proteins that are coded by the nuclear genes which are then to be imported into chloroplast (Martin and Herrman, 1998). As a survey of selected 153 embryophyte plastid genomes, Fig. 1 exhibits the genes present in the chloroplast for the most common gene sets and for the missing genes, and shows the similarity of closest species plastid gene to the nucleus genome of the species.

Despite the dissimilarities in the length of the chloroplast genomes, a typical feature of these genomes is their relatively large inverted repeat (IR) regions. Due to high level of size variation and even its absence in red algae (Rhodophyta), the functional significance of this structure is not fully understood. On the other hand, the slower rate of nucleotide substitution rate of this region compared to the rest of the genome, might shed some light on the promotion of the genome stability hypothesis for this unique genomic structure (Maier *et al*., 1995).

[2] Indeed, the ratio of the transferred genes into the nucleus to the ones remained in the organellar genome, the photosynthesis and respiration genes among other families are having the highest value (close to one). This is followed by the translation and amino acid biosynthesis. Genes related to transport and binding proteins have transferred the most (158 genes, 4 remained in chloroplast and 5 in mitochondrion) to the nucleus while all the genes in central intermediary metabolism are transferred to the nucleus (Martin & Hermann, 1998).

These IR regions inevitably section the genomes into a quadripartite structure consisted of long and small single copy regions (LSC and SSC, respectively), which are separated from each other with two IR regions namely referred to as IRb and IRa (Fig. 2).

Function	Genes	Comments
RNAs		
Ribosomal	***rns***, ***rnl***, ***rrn5***	4.5S rRNA in plants only
Transfer	*trnA(ugc), trnC(gca), trnD(guc), trnE(uuc), trnF(gaa), trnG(gcc), trnG(ucc), trnH(gug), trnI(cau), trnI(gau), trnK(uuu), trnL(caa), trnL(uaa), trnM(cau), trnN(guu), trnP(ugg), trnQ(uug), trnR(acg), trnR(ccg), trnR(ucu), trnS(gcu), trnS(uga), trnT(ugu), trnV(uac), trnW(cca), trnY(gua)*	
Others	*rnpB (ribonuclease P), ffs RNA (SRP), ssra(tmRNA)*	
Transcription	*cbbX, rbcR, rpoA,* ***rpoB***, ***rpoC1***, ***rpoC2***, *matK*	*matK* in greenline, *cbbX and rbcR* in redline
Translation	*tufA*	
Ribosomal proteins		
Small subunit	***rps2***, ***rps3***, ***rps4***, *rps5, rps6,* ***rps7***, ***rps8***, *rps9, rps10,* ***rps11***, ***rps12***, *rps13,* ***rps14***, *rps16, rps17,* ***rps18***, ***rps19***, *rps20*	All plastid in redline
Large subunit	*rpl1,* ***rpl2***, *rpl3, rpl4, rpl5, rpl6, rpl11, rpl12, rpl13,* ***rpl14***, ***rpl16***, *rpl18, rpl19,* ***rpl20***, *rpl21, rpl22, rpl23, rpl24, rpl27, rpl29, rpl31, rpl32, rpl33, rpl34, rpl35,* ***rpl36***	All plastid in redline and in some greens
Photosynthesis		
ATP synthase	***atpA***, ***atpB***, *atpD,* ***atpE***, ***atpF***, *atpG,* ***atpH***, *atpI*	All plastid in redline
Photosystem I	***psaA***, ***psaB***, ***psaC***, *psaD, psaE, psaF,* ***psaI***, ***psaJ***, ***psaL***, ***psaM***	*psaD, E, F* not in plants
Photosystem II	***psbA***, ***psbB***, ***psbC***, ***psbD***, ***psbE***, ***psbF***, ***psbH***, ***psbI***, ***psbJ***, ***psbK***, ***psbL***, ***psbN***, ***psbT***, ***psbV***, ***psbX***, *psbY, psbZ, psb28*	
Cytochrome complex	***petA***, ***petB***, ***petD***, *petF,* ***petG***, *petL(ycf7), petM (ycf31),* ***petN(ycf6)***	*petF* nuclear in plants and many algae
NADH dehydrogenase	*ndhA, ndhB, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*	Plants only, except some gymnosperms
Metabolism	*accD, acpP, chlB,* ***chlI***, *chlL, chlN, rbcL*, rbcS*, thiG, thiS, cysA*	*accD* plants only, **rbcL* and *rbcS* have different origins in red and green line (see text)
Protein quality control	***clpC***, *clpP, dnaB,* ***dnaK***, *ftsH(ycf25),* ***groEL***	
Assembly, membrane insertion	*ccs1, ccsA, secA, secG, secY, sufB, sufC, tatC*	Redline only

Table 1. Plastid genes and their function in plastid genomes. The greenline refers to plants and green algae while redline indicates the red algae and chromistan algae that obtained their plastids from the red algae by secondary endosymbiosis. The bold ones are most prevalent ones (except in reduced genomes like dinoflagellates, non-photosynthetic plastids).

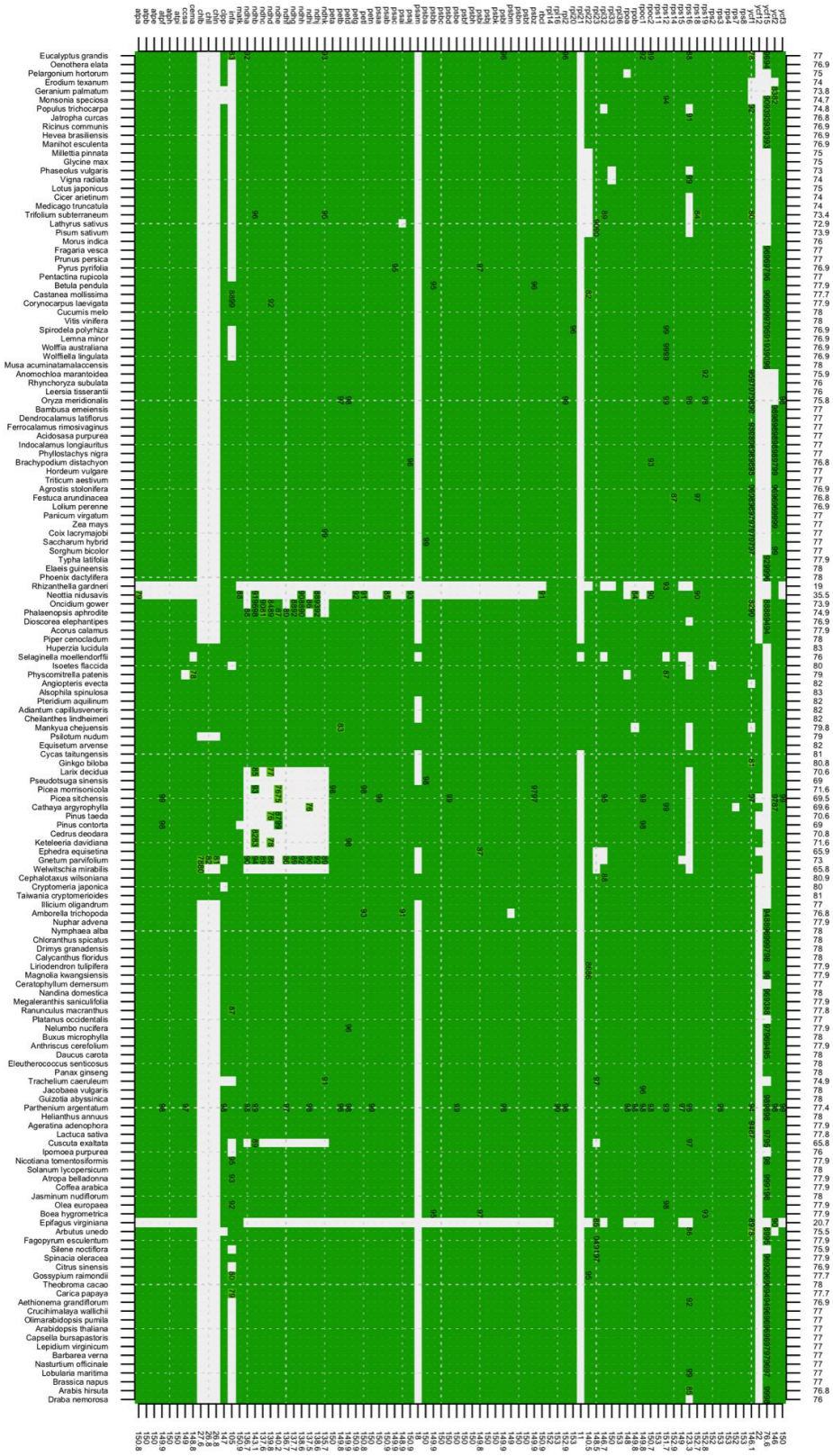

Figure 1. The chloroplast genetic map with nuclear blast hits. The absence/presence diagram of plastid genes in 153 embryophytes. The species are listed on the left with the set of 84 common genes on the top. The corresponding coordinate for a specific species and gene is plotted as either a green or white square for presence or absence of the gene on the plastid genome, respectively. If there is a value plotted in a coordinate on the field it means that the underlying plastid sequence of that species is missing the annotation of that gene and that value represents the blast similarity between that specific plastid gene of the closest neighboring species (target) to the underlying species plastid genome sequence (query). The values plotted on the right and below of the field show the sum of the blast similarity hit percentages plus the present gene and species numbers, respectively.

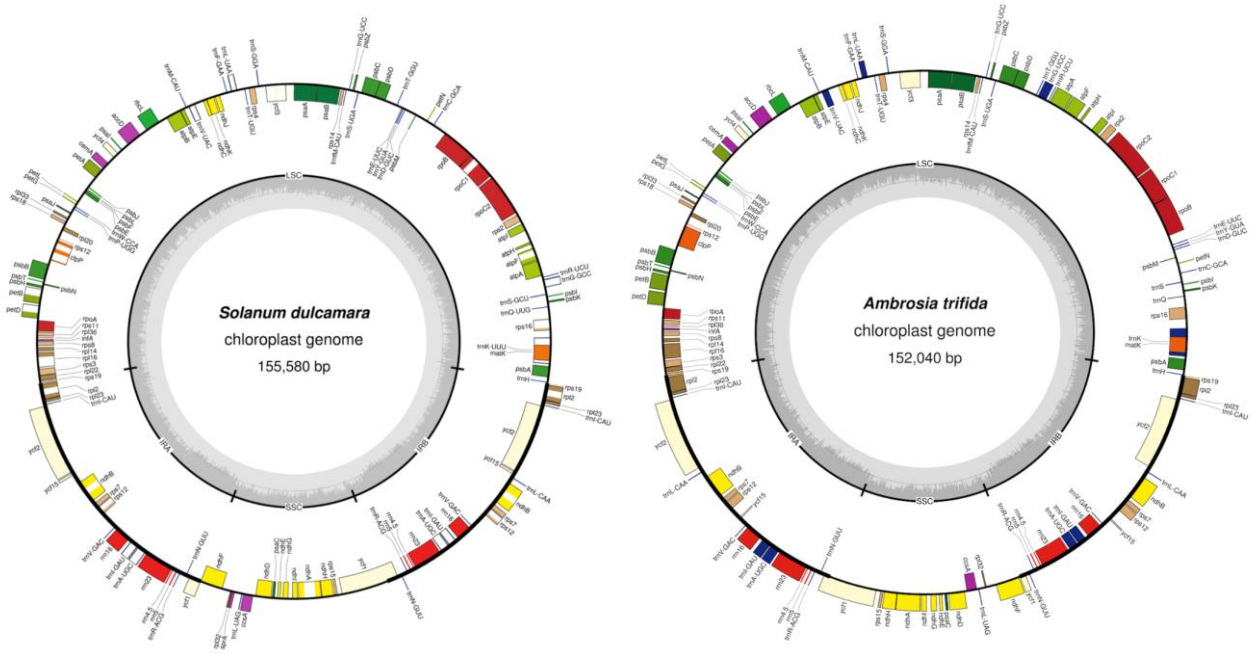

Figure 2. Map of the chloroplast genome of the *Solanum dulcamara* and ***Ambrosia trifida* as depicted in Chapter I and II.** Genes lying inside of the outer circle are transcribed counterclockwise while those outside that circle are transcribed clockwise. Genes belonging to different functional groups are color coded differently and the GC and AT content of the genome are plotted on the inner circle as dark and light gray, respectively. The inverted repeats, large single copy, and small single copy regions are denoted by IR, LSC, and SSC, respectively.

As shown in Chapter **I** and **II**, the chloroplast genome of the *Solanum dulcamara* and *Ambrosia trifida* exhibit this preserved structure as well (Fig. 2). For *S. dulcamara* with the chloroplast genome length of 155,580 bp these regions are 85,901 bp, 18,449 bp, and 25,615 bp for LSC, SSC, and IR respectively. On the other hand, the *A. trifida* corresponding sections length are 83,966 bp, 17,894 bp, and 25,090 bp while the total genome length is 152,040 bp. Chapter **I** further shows that the chloroplast genome of *S. dulcamara* contains 81 protein-coding 27 tRNA, and four rRNA genes that makes the total of 114 unique genes of this genome. Out of these genes, 17 contained introns and among them *ycf3* and *clpP* contained two introns. All these introns are observed to belong to the group II introns while only *trnL-UAA* exhibits the group I intron. The longest forward repeat with size of 83 bp was found in the IGS region of *ycf3* and *trnS-GGA*. Chapter **II** on the other hand reports 80, 28, and four unique protein coding, tRNA, and rRNA genes on the *Ambrosia trifida* chloroplast genome. While overall, these two sampled species, one from the largest genera of angiosperms, and the other from ragweed genus exhibit the standard chloroplast structure, this is not necessarily true for a given plastid genome. Organisms with highly atypical chloroplast genomes are parasitic plants that have lost the ability to carry out photosynthesis (Wolfe *et al*., 1992). Their chloroplasts either have completely lost all the genes that code for photosynthetic proteins, or while still retaining a small vestigial chloroplast genome that functions primarily in fatty acid synthesis, and not in photosynthesis (McFadden *et al*., 1996; Waller *et al*., 1998). One hypothesis put forward for this is underlying the chloroplast degradation process that happens in response to the loss of functional complexes which disrupt the selection pressure rhythm.

This will gradationally transform the organelle into the obligate parasitism (and eventually holoparasitism) state where functional constraints are relaxed (Wicke *et al.*, 2011). The other interesting aspects of plastid genomes is their relatively low AT content. This is prevalent both in the genic and intergenic regions. With roughly 80% genic region of the sequences, the GC content of giant ragweed and bittersweet as presented in Chapter **I** and **II** are observed to be fairly similar (AT content of 37.2% and 37.8%, respectively). These values were somehow close to those of *Nicotiana tabacum* (38%), *Porphyra purpurea* (Roth) C.Agardh (33%), and *Odontella sinensis* (Greville) Grunow (32%; Reith and Munholland, 1995; Kowallik *et al.*, 1995). Although it is not possible to exclude the hypothesis that the low AT of plastids originated from their common cyanobacterial ancestor, it seems more plausible that the endosymbiosis has been the main reason of this (Howe *et al.*, 2002). This drift in nucleotide composition could be due to the nature of the DNA damage occurring in chloroplast, the tendency of the plastid DNA polymerase to mis-incorporate A and T rather than G and C in replications, or a bias in DNA repair machinery (Lang *et al.*, 1999).

Although studying specific characters of a species can be valuable by itself, it would be only in the light of the comparison with other closely related species that the true importance of the information could be embraced. With reference to two families of plants, namely Solanaceae and Asteraceae, next subsection discusses the position of *Solanum dulcamara* and *Ambrosia trifida* in more detail.

2.2 Comparative inference

The comparative genomic analysis accounts for the methods that with a reference to a target sequence, explain the features of the newly sequenced species (Rubin *et al*., 2000). In a crude sense, this can be implied as the supervised machine learning method in which the aim is to estimate the parameters of the newly arrived test data point with reference to the training data. The more erroneous part of this method hence is not to find the similarities between the objects being compared but, the differences that distinguishes each object from the rest and grant them a unique identity. The other important factor that need to be accounted for in the comparative context, is the scope of the set of objects being compared; as this scope being too wide might lead into the intractability of comparisons due to a huge dissimilarity level between objects, while on the other hand, if the scope is too narrow, we fail to notice and infer the unique interesting biological aspects of the sequences.
The eudicot clade of the angiosperms (flowering plants) is the most diverse group of embryophytes (land plants) consisting of 416 families, the Chapter **I** and **II** comparative analysis is focused on two highly important families: Solanaceae (nightshades) and *Asteraceae* (composits). While the nightshades are an economically important family of flowering plants, the composits, competing with Orchidaceae, is arguably the biggest family with 32,913 species and 1,911 genera (Stevens, 2001).

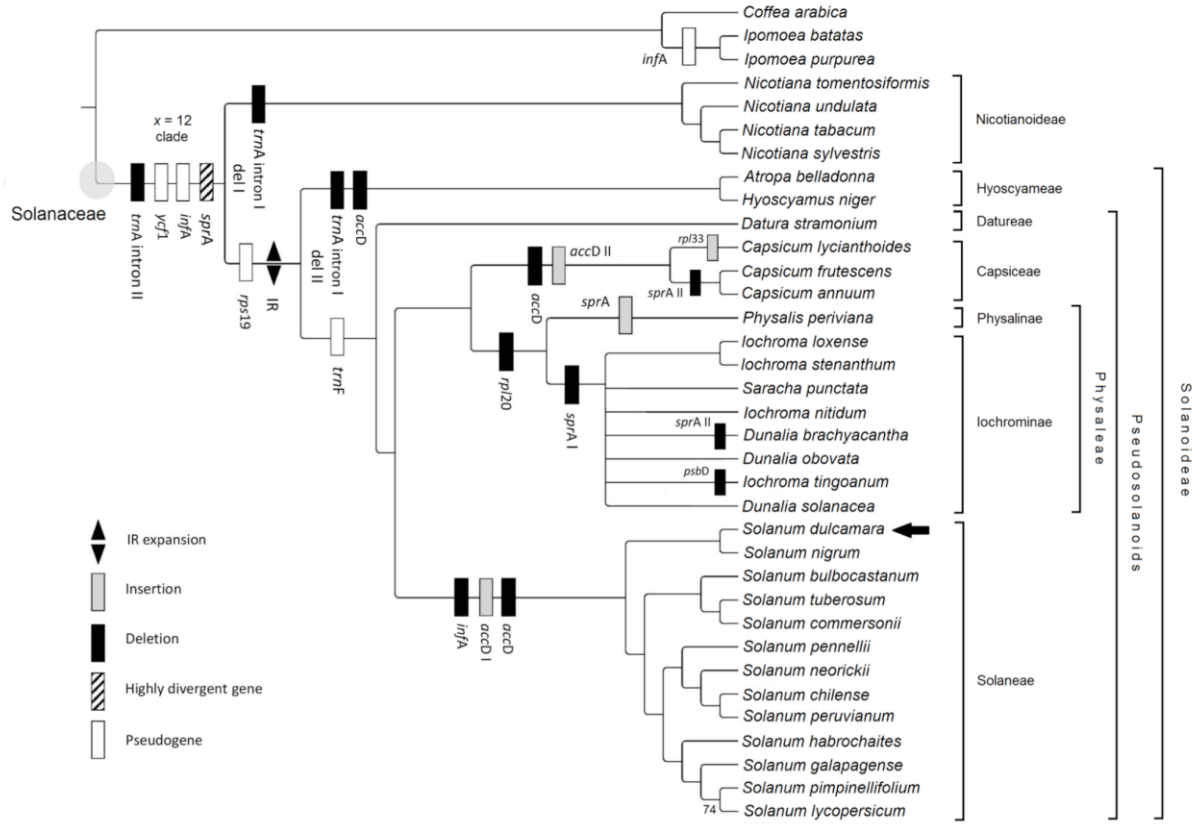

Figure 3. Cladogram illustrating the phylogenetic relationships of Solanaceae based on complete chloroplast genome sequences. Plastid genome rearrangement events are mapped on the branches of the best scoring maximum likelihood tree generated with RAxML-NG. Each node has 100% bootstrap support value. A node with lower support value indicated and those with support values below 50% collapsed. Currently recognized suprageneric groups are listed on the right.

In Chapter **I**, based on the whole genome alignment, we presented a phylogenetic tree of 32 chloroplast genomes of Solanaceae and ran a survey of the evolutionary events pertinent to this species set with *Coffea arabica* L., *Ipomoea batatas* (L.) Lam., and *I. purpurea* L. as outgroup terminals (Fig. 3). MAFFT (Katoh, 2013) was used to align the 35 complete chloroplast genomes and maximum likelihood (ML) analyses were performed with RAxML-NG (Kozlov, 2018). Three strategies in regard to the alignment was used. 1) The exclusion of one of the IR regions to reduce overrepresentation of duplicated sequences and then treating the unpartitioned alignment under GTR+I+G substitution model as a single partition; 2) Partitioning the same data matrix by gene, exon, intron and intergenic spacer regions (n = 258) and allowing separate base frequencies, α -shape parameters, and evolutionary rates to be estimated for each; 3) Using the PartitionFinder2 (Lanfear, 2017) to infer the best-fitting partitioning strategy for the alignment (n = 24). jModelTest2 (Darriba, 2012) was used to infer the best fitting nucleotide substitution model, and the branch support values were obtained by 10,000 rounds of bootstrap. For each alignment we conducted ten separate runs

with RAxML-NG since log-likelihoods could show variation among individual runs (Nguyen, 2015). This phylogenetic analysis resulted in highly resolved tree with almost all clades recovered with maximum branch support values.

We further compared the gene order, inverted repeat (IR) length and studied the gradational structural changes in the family. This would have not been possible with the quality of the existing annotations as there are large number of errors in the deposited annotations. In order to have a higher resolution comparison, we first revised and reannotated the plastid genomes of the Solanaceae. The process involved two steps. We first used the predictive tools to detect the genes and then manually annotated all the genes of the genomes. For example, we noticed that the *Iochroma loxense* (Kunth) Miers and *Solanum pennellii* Correll genome sequences completely lacked the annotation and genomics features. In general, these annotation errors could cause a considerable difficulties and downstream errors in inferring the gene functionality, synteny and even in phylogenetic analyses.

It turned out that the ancestral chloroplast genome of *Solanaceae* most likely had two pseudogenes: *infA* and *ycf1*. *ycf1*a and *ycf*b loci were detected to be the most diverse regions of the chloroplast genomes and the former has been suggested as a "barcode" for embryophytes (Dong *et al.*, 2015). The frequent and multiple parallel losses of *infA* gene to the nucleus and its defunctionalization distinguish this gene as the most mobile chloroplast gene known in plants (Millen *et al.*, 2001). Other interesting observation regarding our sequences is the highly divergent *sprA* (Fig. 3). The existence of this gene is nonessential in maturation of the pre-16S rRNA in plastids (Sugita *et al.*, 1997). Furthermore, while the distribution of the genes of different regions of the genome resembles other *Solanaceae* (with 13 genes on SSC and 19 genes on the IR) our ancestral genome reconstruction analysis suggested the further rearrangement and expansion of the IR in the common ancestor of *Solanoideae*. This expansion (ranging from 25,343 bp to 25,906 bp), was further assessed with the examination of the junction sites of the curated genome annotation of nine other Solanaceae (Fig. 4).

Following similar methodological approach, we studied the chloroplast genome of the *Ambrosia trifida* in Chapter **II**. The phylogenetic analysis is based on 41 different genomes from the Asteraceae with the *Carum carvi* L. and *Foeniculum vulgare* Mill. as the outgroup terminals. For this, we used the matrix based on 50 protein-coding genes representing 43 species resulting in a total concatenated matrix alignment of 31,356 bp. While it has been shown that the non-coding regions in chloroplast can be informative to infer the phylogenies (Kelchner, 2000), in this case, due to the observed ambiguity in the repeat expansion and presence of microstructural variation, these regions were excluded (Curci *et al.*, 2015). Our results are congruent with the current hypotheses about phylogeny of Asteraceae. The placement of the genes in the vicinity of the IR junction sites was also analyzed with the

species of the Asteraceae included in the phylogenetic analysis. This exhibited the placement of the *rps19* in the LSC-IRb region. Majority of the species showed the fixation of the extended *ycf1* on the JSB and length of the IRs, SSC, and LSC was quite conserved throughout all the species as 24-26 kbp, 18-19 kbp, and 82-84 kbp, respectively. In general, the overall synteny of the genes in the vicinity of the junction sites was confirmed as well.

As discussed earlier the comparative task can entail its own challenges which renders the true comparative analysis unattainable; nevertheless, a close examination of subjects can at least deliver clues on which direction the next logical question shall be placed: a) One interesting case is the exceptionally short SSC region (~2 kbp) of the Vanilloideae (Orchidaceae; Ueda *et al*., 2012). This short SSC is the result of the multiple loss and pseudogenization of the *ndh* genes of the chloroplast. In *Vanillon* for example, only *ndh*B was detected in the plastid genome while all other 10 plastids encoded NAD(P)H dehydrogenase complex genes have either disappeared, or are non-functional (Amiryousefi *et al., 2017*). In general, this independent and complex rearrangement pattern of genes in Orchidaceae is not observed to be tied with a meaningful evolutionary event (Kim *et al*., 2015). b) Another interesting feature is that the IR regions cannot be identified. For example, *Guizotia abyssinica* L. plastid genome showed the dissimilar IR regions as these two regions were interspersed by single nucleotide and insertion-deletion polymorphism (Chapter **I**). While eight SNPs and a deletion of 2 bp from the IRb may seem minuscule against the 25,001 bp long IR, they are significant enough to cause the incorrect detection of these regions (Dempewolf *et al*., 2010). Although it is tempting to assume this difficulty to identify these regions is inherent to the IRs, the sequencing and/or assembly error hypothesis seems to be more likely of an explanation. In either of the two examples given above, hence it is crucial to obtain the highest level of assurance about our measures. While the case with Vanilloideae, deserves an evolutionary endeavor to answer this family's peculiarity, the curation of the plastid genome annotation is most probably the remedy for the *G. abyssinica* case.

As mentioned earlier, the type of the errors that might occur in the comparative inference can be either misinterpretation and inferring a type of difference that is not present in the target sequence, or, on the other hand, fail to detect a similarity that is present in the sequences studied and compared. Both of these cases can be observed for example in the annotation tools commonly used, more specifically in a highly cited program for organellar genome annotation: DOGMA (Wayman *et al.*, 2004). Our surveys of Solanaceae presented in Chapter **I** for example revealed a huge deflection from correct annotation of the species in this family. What we have observed is probably by no means exclusive to this family. These types of errors continue to linger on in any type of inferences as long as the critical quality checks are not routinely included in handling the obtained results produced using the tool mentioned.

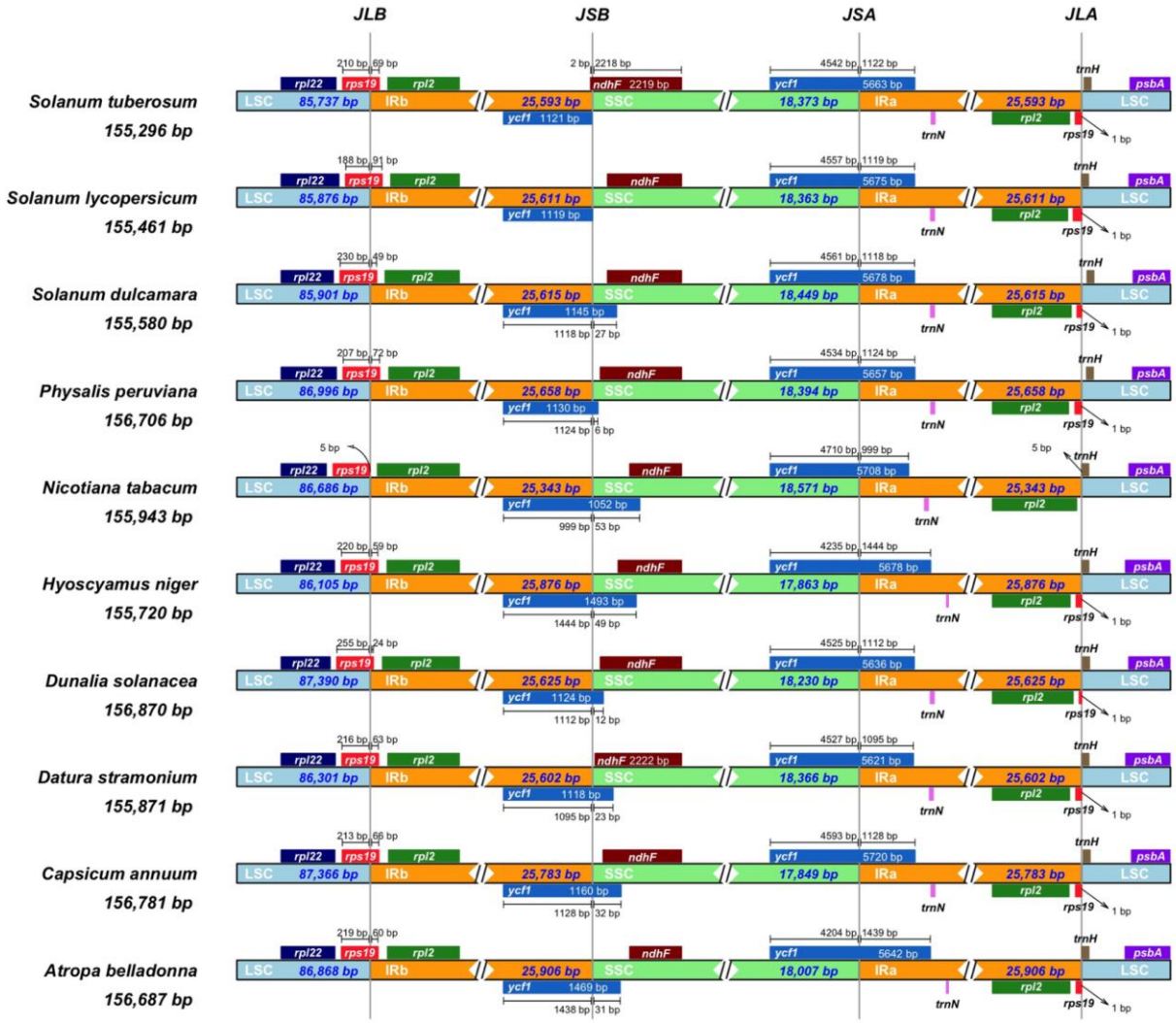

Figure 4. Junction sites of the inverted repeats. For each species, genes transcribed in positive strand are depicted on the top of their corresponding track with right to left direction, while the genes on the negative strand are depicted below from left to right. The arrows are showing the distance of the start or end coordinate of a given gene from the corresponding junction site. For the genes extending from a region to another, the T bar above or below them show the extent of their parts with their corresponding values in base pair while nothing is plotted for the gene tangent to the sites. The plotted genes and distances in the vicinity of the junction sites are the scaled projection of the genome. JLB (IRb /LSC), JSB (IRb/SSC), JSA (SSC/IRa) and JLA (IRa/LSC) denote the junction sites between each corresponding two regions on the genome.

Here with reference to two chloroplast genomes, we presented the basics of the structure and architecture of the plastid genomes. Stepping up, in comparison to the other species, we discussed ways we can improve quality of the deposited chloroplast genomes and what can be inferred from comparative analyses. In order to improve our fourth type of downstream inference, the next section introduces two software packages. More specifically, as respectively published in Chapter **III** and **IV**, a tool to evaluate the chloroplast genomes in the vicinity of the junction sites and, a calculative online tool to ease the process of obtaining seven indices related to primary markers are discussed.

3. New age informatics

"From where I stand, the rain hit the ground at random, if I could stand somewhere else, I could see the order in it."

-Tony Hillerman

The electron microscope showed us a lot about the nanoworld but our knowledge about that realm would have not improved if it was not about our better apprehension of the realities beheld by our eyes. While the rate of our recognition of the nature has always been proportional with the invention and enhancements of the new tools, it seems that we are now approaching the time when we are about to be left behind the supersonic wheel of technological advancements and flood of novel data. This is indicating an inevitable gap - a gap between our ability to internalize and process the information with the tools we have created. Specifically, regarding the genomics, there are ample methods that can be used in inferring different queries ranging from phylogeny to the structure and functions. The ample number of tools, however, should not be considered synonymous with good quality. As shown in Chapter **I,** there were many errors represented in the plastid annotations of the Solanaceae which we corrected to render our downstream analysis possible. One of the most problematic area of existing annotations was related to the standard quadripartite structure of the angiosperm plastid genomes. In number of cases we noticed that LSC and SSC were entirely missing or poorly indicated. Inverted repeats (IRs) were either unannotated or their orientation, size and correct naming was erroneous. Compared to the tobacco reference order LSC-IRB-SSC-IRA (Shinozaki *et al.*, 1986), the erroneous annotation LSC-IRA-SSC-IRB was often implied. These discrepancies in the annotations could have many different reasons out of which the inefficiency of automated analytical tools can readily be recognized.

In Chapter **III** we present a tool to detect this structure in the chloroplast and visualize these areas. This new tool can visualize the junction sites of the chloroplast genome of up to ten different embryophytes. The report of this sort in the comparative sections in the studies of plastids is an indispensable part but previously such a tool towards this end did not exist. Previously the plots were compared only pairwise and they were of poor quality (Terakami *et al.*, 2012; Pląder *et al.*, 2007; Li *et al*., 2013). This tool is coded in R and it is available for online use as well as a source code. The program allows direct submission of files in either GB format or insertion of the GI or accession number of the selected species. It is possible to input the data also manually in the DOGMA (Wayman *et al.*, 2004) file format. The software preprocesses the input files in various ways, and after success, enters into the IR finding stage. As mentioned in Chapter **I**, SNPs on the IR region will not pose serious difficulties for the

program to detect these areas. For a given set of species, the program finds the consensus radius for each junction sites as the best visual representation of the genes. Also, in the program is embedded an SSC reversion option that allows the reverse representation of this region and its pertaining genetic annotations. This deemed to be essential as chloroplast DNA within individual plants may exist in two equimolar states that differ in the relative orientation of the SSC region (Palmer 1983; Walker *et al*., 2015). Upon successful submission of the input files and consecutive calculations, the program delivers a high-quality jpg-file with the depicted tracks of different species and their detailed genetic content at the base pair indicating resolution.

Some of the existing genomic plotting have targeted more specifically for synteny analysis (Lyons and Freeling, 2008), comparative visualization (Frazer *et al.*, 2004), or genome organization plot (Lohse *et al.*, 2007), and are not necessarily presenting such a high-resolution analysis and output presented in Chapter **III** which is optimally designed for such a task. This is because of unique and compact nature of the plastid sequence; the composition of the plastomes with relatively limited number of functional genes leaves only a fraction of the genome for intergenic regions (as also observed in Chapter **I**, **II** this proportion was only 20% for the bittersweet and giant ragweed). Another interesting feature of the plastid genomes that renders the existing plotting tools unscalable is the existence of the two stretched IR regions of 15-32 kb (Oldenburg and Bendich, 2004). These regions are attached within the LSC and SSC regions on four distinct JS (Fig. 4). The structural organization of the genome along these sites are of crucial importance, examination of which, can reveal sweep or evolutionary drift in lineages. Furthermore, the interconversion into a dumbbell-shaped conformation of circular plastid molecule, is facilitated by the IRs (Kolodner *et al.*, 1976). Also, the contraction of the IR and SSC is primary reason for the variation in size of angiosperms plastid genomes. The presented tool in Chapter **III** has been tested on the diverse set of more than 200 species, and seems to show well the discussed specifics of the plastid genomes.

The other interesting question to be answered with bioinformatics is related to the characterization and evaluation of genetic diversity within and between species and populations. One type of the valuable tools available today are the molecular markers that have proved to be useful in this context. Different markers disagree in their pertinent information content, which is mostly dependent on polymorphism (Nagy *et al.*, 2012). Defining the genetic variation in a population, the concept of polymorphism, is used in diverse fields and disciplines ranging from genetics, microbiology, conservation biology, to botany and zoology (Mukherjee *et al.*, 2010; Muneer *et al*., 2011; Rajkumar *et al*., 2011). The ample use of the methods emphasizes the important basic questions involved. In essence, they can be considered to be centered around the difficulty of finding the useful polymorphic loci, the number of markers needed, etc. These concerned issues related to the markers can be tackled with measuring their information content. Different indices have existed for a long time

(Botstein *et al.*, 1980), and attempts has been made to make them easily accessible (Nagy *et al.*, 2012). However, there have not been a comprehensive online tool to collectively assess these indices.

As summarized in Table 2., Chapter **IV,** provides an online platform for calculating seven different marker efficiency indices, namely: heterozygosity index (*H*), polymorphic information content (*PIC*), effective multiplex ratio (*E*), discriminating power (*D*), marker index (*MI*), arithmetic mean heterozygosity (H_{avp}), and resolving power (*R*). The indices are based on dominant and codominant DNA fingerprinting markers. Calculating these indices allows comparison and selection of optimal genetic markers for a given data set. The platform, called iMEC, is as well, completely written using R and is publicly available. The input file format for the program can be either PHYLIP (Felsenstein 2002), NEXUS (Maddison *et al*., 2017), or a simple excel file. The input data should be either binary coded or recorded as multistate characters, both cases represented in integer form.

Index	Formula	Definition
Expected heterozygosity[a]	$H = 1 - \Sigma p_i^2$	The probability that an individual is heterozygous for the locus in the population. p_i is the allele frequency for the *i*-th allele, and the summation is over all available alleles.
Polymorphism information content[b]	$PIC = 1 - \Sigma p_i^2 - \Sigma\Sigma p_i^2 p_j^2$	The probability that the marker genotype of a given offspring will allow deduction, in the absence of crossing over, of which of the two marker alleles of the affected parents it received. p_i and p_j are the population frequency of the *i*-th and *j*-th allele. The first summation is over the total number of alleles, whereas the two subsequent summations denote all the *i* and *j* where $i \neq j$.
Effective multiplex ratio[c]	$E = n\beta$	The product of the fraction of polymorphic loci for an individual assay. In other words, the number of loci polymorphic in the germplasm set of interest analyzed per experiment fraction of polymorphic loci. Defining $\beta = n_p / (n_p + n_{np})$, where *p* and *np* indicate the polymorphic and nonpolymorphic fraction of the markers, so n_p and n_{np} represent their respective counting numbers.
Mean heterozygosity[c]	$H_{avp} = \Sigma H_n / n_p$	The average heterozygosity calculated for polymorphic markers. H_n is the heterozygosity of the polymorphic fraction of markers, and the summation is over all of the polymorphic loci n_p.
Marker index[c]	$MI = E H_{avp}$	The product of the effective multiplex ratio and the average expected heterozygosity for polymorphic markers, where H_{avp} denotes the average expected heterozygosity for the polymorphic markers. It is equal to $\Sigma H_p / n_p$, where the summation is over all polymorphic sites with H_p and n_p defined as above.
Discriminating power[d]	$D = 1 - C$	The probability that two randomly chosen individuals exhibit different banding patterns and are thus distinguishable from one another. *C* is defined as the confusion probability. For the *i*-th pattern of the given *j*-th primer, present at frequency p_i in a set of varieties, the confusion probability is $C = \Sigma c_i = \Sigma p_i \frac{Np_i - 1}{N-1}$ where for *N* individuals, *C* is equal to the sum of all c_i for all of the patterns generated by the primer.
Resolving power[e]	$R = \Sigma I_b$	Resolving power is based on the distribution of alleles within the sampled genotypes and strongly correlates with the ability to distinguish between analyzed samples. The division of samples into two groups is based on the presence or absence of a band, ideally present in one part of the samples while absent from the other. Bands can be weighed according to their similarity to the optimal condition (50% of genotypes containing the band), where I_b or band informativeness is represented on a scale of 0–1 and is defined as $I_b = 1 - (2 \times \lvert 0.5 - p \rvert)$, where *p* is the portion of the samples containing the observed band. Using this value, the resolving power or the ability of a primer (technique) to distinguish between genotypes could be represented by the sum of these adjusted values for all generated bands.

Table 2. Detailed description of polymorphism indices calculated by iMEC. For each marker as noted with the superscripts the references are as follow. [a] (Liu, 1998); [b] (Botstein *et al.*, 1980); [c] (Powell *et al.*, 1996); [d] (Tessier *et al.*, 1999); [e] (Prevost and Wilkinson, 1999).

Note that the inherent complexity of the evolution plays a major role in the shortcomings of the many analytical tools. For example, endosymbiotic gene transfer and replacements are such phenomena related to the plastids (Lane *et al.*, 2008). But what is left after that in the detection of the genes in the chloroplast genome, is the incompetence of the either sequencing, assembly, and/or annotation procedures. The latter issue as discussed in detail in Chapter **I**, and is particularly important to be addressed as all the other downstream biological analysis

are expected to be built upon that. The other danger of poor annotation is that mistakes can easily be transmitted to a newly sequenced genome if that is used as the annotation model. One such sweep in annotation quality observed in Solanaceae was caused with a widely used annotation tool: DOGMA (Wyman *et al.*, 2004). This situation can be even exacerbated due to the fact that this program generates the general feature format (.gff) and GenBank (.gb) output files that can be easily incorporated in other software and errors propagated even further. Luckily new programs like CpGAVAS (Liu *et al.*, 2012) and GeSeq (Tillich *et al.*, 2017) have been introduced that can perform better and seem to become increasingly popular. On the other hand, the use of a complementary tools to benchmark different annotations such as BEACON (Kallkatawi *et al.*, 2015) seems to become prevalent.

The other dimension of evolutionary studies is the use of informatics in resolving the phylogenetic relationship between set of taxa. Due to its very nature there is perhaps not a more controversial realm in evolutionary studies than this one. Some subjectivity is unavoidable, although rarely discussed in detail. This ranges from the choice of substitution models, to the metric measures for the distance matrix, to the methods of deducing the trees, and even the assessment of the obtained topology. This subjectivity is one of the reasons for the lack of consensus between scientists. For example, despite the lack of evident theoretical justification (Holmes, 2003) the bootstrap for assessing the phylogenies (Felsenstein, 1983) is still widely used, and for some scientists this seems to be even more important than the actual result(s) obtained using the chosen optimality criteria. Also, whether to use all information obtained from the genomes, or only part of it, has been discussed (Salichos and Rokas, 2013), where they have suggested using the genes with strong evolutionary signals. As well in a sequel paper (Salichos *et al.*, 2014) they propose a new entropy-based index that calculate a balanced score with incorporating the number of the gene tree topologies favoring or against the derived species tree named as internode certainty. Based on this score, and albeit in contrary to the widespread practices (Zhang *et al.*, 2012; Lang *et al.*, 2013), they argue that the use of the slowly evolving genes and conserved sites increases the incongruency (i.e. decreases the certainty) of many internodes of the derived phylogeny. Hence, they recommend the detection and use of the genes with strong evolutionary signals for constructing the robust phylogenetic tree. This recommendation has been challenged with the objection that it is not the evolutionary signals of the gene that plays role in the internode certainty, and that their derivations is merely a specific gene selection artifact (Betancur-R *et al.*, 2014). This refuting argument is based on the illustration of the significant negative correlation between gene length and their level of incongruence, as such relationship has already been established (Rasmussen & Kellis, 2007). This is because the shorter genes mean smaller nucleotide sample size (and hence a more conflicting gene-trees) and it is possible to obtain all conflicting trees with these genes (Dikow & Smith, 2013).

4. Conclusion

"Improvement makes straight roads, but the crooked roads without improvement, are roads of genius."

- William Blake

The progression of science and its impersonality has never been felt more than now. The instant gradational increase of our knowledge is bestowed both in the invention of modern tools and production of more accurate hypotheses. More specifically, in the molecular biology this can be translated in the accumulation of the sequenced genomes of different species which consequently can be used to present better hypotheses about many pertinent aspects of the organisms; from their ontogeny to their interaction with the surrounding environment and to hypotheses about their phylogeny.

This study is based on two main parts that are inevitably related to each other. Firstly, we were interested about plastid genomic material of two angiosperms and then addressed suggestive downstream questions. Using plastid genomes of the closely related species we examined phylogeny of these two species. We detected different genetic events that characterize species studied. We also curated the erroneous plastid genome annotations of the Solanaceae in Chapter **I**. The analyses performed were, however, by no means exhaustive and the material available in GenBank should be used for performing new analyses and testing different hypotheses presented. Secondly, we focused on the development of two separate tools to perform analyses of the material obtained for example from the first part. These tools are embedded as online interactive suits that are meant to ease the representation of the various evolutionary aspects of the input data. In chapter **III** for example we designed a tool to depict the genetic content of the chloroplast genomes of embryophytes, while Chapter **IV** represents a platform to calculate different molecular indices of informativeness.
While the effort and focus of this thesis is mainly placed on plastids and their genomic content, one should be aware of the shortcomings of the analyses and outlook. There are many other valuable sources of information that are useful for comprehensive analyses. The immediate genetic counterparts of the plastids – those of the nucleus and mitochondrion should be used whenever available in order to obtain robust hypotheses.

Besides, we need to bear in mind the limits of the extent of our methods and more than that, our ability to discern forces of nature as reflected in the torturous evolutionary history. We may lure into thinking that nature follows a certain model, but that might not be necessarily true, as ways of nature can be grander that our collective consensus. This should not stun us however, as it has formed us to inexorably seek the truth.

ACKNOWLEDGMENT

First, I would like to express my utmost gratitude to my supervisors Prof. Jaakko Hyvönen and Dr. Péter Poczai who were there anytime I needed them and their guidelines were essential for this to happen. Thank you for letting me to have the opportunity to be your student and enjoy this fun and fruitful milestone by your sides. Specifically, thank you Péter for boosting my self-confidence with teaching me how to grasp and appreciate my little ideas and Jaakko for showing me what is the meaning of a kind and caring supervisor. I am thankful to my supportive thesis advisory committee members Prof. Teemu Teeri and Prof. Eva-Mari Aro for finding time to meet and discuss my work and appreciate their help and suggestions which guaranteed my smooth progress. I like to thank Prof. Jaakko Kangasjärvi for granting me the opportunity to initiate my studies in plant stress group and enjoy being part of a big and cooperative community and also like to thank Dr. Jarkko Salojärvi for trusting me to pursue my questions in a completely different realm of studies than my background. I also like to thank all the friends and colleagues of the CoE for Molecular Biology community as well as Viikki plant science division for forming a lively and dynamic atmosphere that positively helped me through my journey. Specially I learned from Prof. Ykä Helariuta, Dr. Alexey Shapiguzov, Dr. Kirk Overmyer, Dr. Peter Gollan, Dr. Ari Pekka Mähönen, Dr. Fuqiang Cui, and Dr. Timo Sipilä. Another small part of this big community which I, in many ways, see myself in debt for these years for their friendship and sincerely more than their positive impact toward my scientific maturity are Omid, who, with his positive outlook besides his hardworking attitude taught a lot to me, Sitaram, Ondrej, Juan, Aleksia, Jorma, Adrien, Kirti, Julia K, Johanna, Riikka, and Riccardo, Juha and Kaisa. Thank you all for giving me beautiful, life lasting memories. Besides, I would like to thank Dr. Karen Sims-Huopaniemi whom never excused me for prompt answering of my frequent questions, Prof. Johannes Enroth for his kind concern toward my graduation matters and also, I am thankful to Prof. Kurt Fagerstedt for his seamlessly kind concern over my bureaucratic issues. I also feel obliged to thank the colleagues and staff of the Finnish Museum of National History, who did their best to ensure my smoothest placement in this unit. Specifically Dr. Marko Hyvärinen and Harri Sihvonen. Finally, I appreciate Prof. Ildikó Karsai for accepting to act as my opponent, and Prof. Françoise Budar and Dr. Endre Barta as the reviewers of this work for their improving comments. Also, thank you Prof. Jouko Rikkinen for acting as my custos.

Last but not least, I need to appreciate my friends and family who their constant help and support surpasses description. They have weaved in me more than I can see; the webs of care: Abbas, Ali, Amin, Arash, Atta, Behrang, Behrooz, Davood, Hadi, Hajar, Hamid, Hugo, Jalal, Jamal, Jukka, Jussi, Kamal, Kourosh, Mahsa, Mansour, Maria, Mehdi, Miika, Neda, Ninni, Saeed, Sanaz, Sasan, Siamak, Soroush, Terry, Parsa, Vahid, Yahya, thank you all to let me grow with sharing with me. Especially Mahnaz and Hossein who never let my happiness to be superseded with any; those whom I belong more than I can imagine.

References

Allen JF. 2015. Why chloroplasts and mitochondria retain their own genomes and genetic systems: Colocation for redox regulation of gene expression. PNAS. 112, 10231-10238.

Allen JF. 1993. Control of gene expression by redox potential and the requirement for the chloroplast and mitochondrial genomes. J. Theor. Biol. 165, 609-631.

Amiryousefi A, Hyvönen J, Poczai P. 2017. The plastid genome of Vanillon (Vanilla pompona, Orchidaceae). Mitochondrial DNA Part B. 2, 689-691.

Archibald JM. 2009. The puzzle of plastid evolution. Curr Biol. 19, R81-R88.

Bellot S, Renner SS. 2016. The plastomes of two species in the Endoparasite genus *Pilostyles* (Apodanthaceae) each retain just five or six possibly functional genes. Genome Biol Evol. 8,1, 189-201.

Betancur-R R, Naylor GJP, Orti G. 2014. Conserved genes, sampling error, and phylogenetic inference. Syst. Biol. 63, 2, 257–262.

Betts HC, Puttick MN, Clark JW, William TA, Donoghue PCJ, Pisani D. 2018. Integrated genome and fossil evidence illuminates life's early evolution and eukaryote origin. Nat Eco Evol. 2, 1556-1562.

Bhattacharya D. *et al.* 2007. How do endosymbionts become organelles? Understanding early events in plastid evolutions. BioEssays. 29, 1239-1246.

Blankenship RE. 2009. Molecular mechanism of photosynthesis. Blackwell publishing.

Blankenship RE. 2010. Early evolution of photosynthesis. Future perspective of plant biology. 154, 434-438.

Blank CE, Sanchez-Baracaldo P. 2010. Timing of morphological and ecological innovations in the cyanobacteria--a key to understanding the rise in atmospheric oxygen. Geobiology. 8,1, 1-23.

Botstein D, White RL, Skolnick M, Davis RW. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. American Journal of Human Genetics. 32, 314–331.

Chen F, Dong W, Zhang G, Guo X, Chen J, Wang Z, Lin Z, Tang H, Zhang L. 2018. The sequenced angiosperms genomes and genomes databases. Frontiers in plant science 9:1418.

Cheung AY, Bogorad L, Van Montagu M, Schell J. 1988. Relocating the gene for herbicide tolerance: a chloroplast gene is converted into a nuclear gene. Proc. Natl Acad. Sci. USA. 85, 391-395.

Criscuolo A, Gribaldo S. 2011. Large-scale phylogenomic analyses indicate a deep origin of primary plastids within cyanobacteria, *MBE*. 28, 11, 3019–3032.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and high-performancecomputing. Nature Math. 9:772.

Dempewolf H, *et al.* 2010. Establishing genomic tools and resources for *Guizotia abyssinica* (L.f.) Cass – the development of a library of expressed sequence tags, microsatelitte loci, and the sequencing of its chloroplast genome. Molecular ecology. 10, 1048-1058.

Deschamps P, *et al.* 2008. Metabolic symbiosis and the birth of plant kingdom. Mol. Bio. Evol. 25, 536-548.

Dikow RB, Smith WL, 2013. Complete genome sequences provide a case study for the evaluation of gene-tree thinking. Cladistics. 29, 6, 672-682.

Dong W, Xu C, Li C, Sun J, Zuo Y, Shi S, Cheng T, Guo J, Zhou S. 2015. *ycf1*, the most promising plastid DNA barcode of land plants. Scientific Reports, 5, 8348-8353.

Douglas, S. E, 1998. Plastid evolution: origins, diversity, trends. Curr. Opin. Genet. Dev. 8, 655-661.

Falcon LI, Magallon S, Castillo A. 2010. Dating the cynobacterial ancestor of the chloroplast. 2010. ISME J. 4, 777-783.

Felsenstein J. 2002. PHYLIP (Phylogeny Inference Package) version 3.6.

Felsenstein J. 1983. Statistical inference in phylogenies. 1983. J.R. Statist. Soc. A. 146, 3. 246-272.

Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. 2004. VISTA: computational tools for comparative genomics. Nucl Aci Res. 32, 273-279.

Green BR. 2011. Chloroplast genomes of photosynthetic eukaryotes. Plant Journal. 66, 34-44.

Holmes S. 2003. Bootstrapping phylogenetic trees: theory and methods. Statistical Science. 18, 2. 241-255.

Howe CJ, Barbrook AC, Loumandou VL, Nisbet RER, Symington HA, Wightman TF, 2002. Evolution of the chloroplast genome, The Royal Society. 358, 99-107.

Hug LA. *et al*. 2016. A new view of the tree of life. Nature Microbiology. 1, 16048.

Joppa LN, Roberts DL, Pimm SL. 2010. How many species of flowering plants are there? Proc. R. Soc. B. 272, 285-287.

Kallkatawi M, Alam I, Bajic VB. 2015. BEACON: automated tool for bacterial genome annotation comparison. BMC Genomics. 16:616.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in performance and usability. Mol Biol Evol. 30, 772-780.

Keeling PJ. 2010. The endosymbiotic origin, diversification, and fate of plastids. Phil. Trans. R. Soc. B. 365, 729-748.

Kelchner SA, 2000. The evolution of non-coding chloroplast DNA and its application in plant systematics. Ann. Missouri. Bot. Gard. 87, 482-498.

Kim HT, Kim JS, Moore MJ, Neubig KM, Williams NH, Whitten WM, Kim JH. 2015. Seven new complete plastome sequences reveal rampant independent loss of the ndh gene family across Orchids and associated instability of the inverted repeat/small single-copy region boundaries. Plos One 10e, 0142215.

Kiureghian AD, Ditlevsen O. 2009. Aleatory or epistemic? Does it really matter? Struct Safety. 31, 105-112.

Kolodner R, Tewari KK, Warner RC. 1976. Physical studies on the size and structure of the covalently closed circular chloroplast DNA from higher plants. Biochim. Biophys. Acta. 447,2, 144-155.

Kowallik KV, Stoebe B, Schaffran I, Kroth-Pancic P, Freier U. 1995. The chloroplast genome of chlorophyll a + c-containing algae, *Odontella sinensis*, Platn. Mol. Biol. Rep. 13, 336-342

Kozlov A, Morel B, Redelings B. 2018. RAxML-NG v0.7.0 BETA (version 0.7.0). Zenodo. http://doi.org/10.5281/zenodo.1469095.

Lane CE, Archibald JM. 2008. The eukaryotic tree of life: endosymbiosis takes its TOL. The Cell. 23, 268-275.

Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. 2017. ParitionFinder2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Mol Biol Evol. 34, 772-773.

Lang JM, Darling AE, Eisen JA. 2013. Phylogeny of bacterial and archaeal genomes using conserved genes: supertrees and supermatrices. PLoS ONE 8, e62510.

Lang BF, Gray MW, Burger G. 1999. Mitochondrial genome evolution and the origin of eukaryotes. A. Rev. Genet. 33, 351-397.

Larkum AWD, *et al.* 2007. Shopping for plastids. Trends Plant Sci. 12, 189-195

Li X, Gao H, Wang Y, Song J, Henry R, *et al.* .2013. Complete chloroplast genome sequence of Magnolia grandiflora and comparative analysis with related species. Sci China, Life Sciences. 56,2. 189-198.

Liu C, Shi L, Chen H, Zhang J, Lin X, Guan X. 2012. CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. BMC Genomics. 13, 715.

Lohse M, Drechsel O, Bock R. 2007. OrganellarGenomeDRAW (OGDRAW) – a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Curr Genet. 52, 267-274.

Lopez-Juez E, Pyke KA. 2005. Plastids unleashed: their development and their integration in plant development. Int. J. Dev. Biol. 49. 557-577.

Lyons E, Freeling M, 2008. How to usefully compare homologous plant genes and chromosomes as DNA sequences. The Plant Journal. 53,4. 661-673.

Nagy S, Poczai P, Cernak I, Gorji AM, Hegedus G, Taller J. 2012. PICcalc: An online program to calculate polymorphic information content for molecular genetic studies. Biochemical Genetics. 50, 670–672.

Nakayama T, Archibald JM. 2012. Evolving a photosynthetic organelle. BMC Biology 10, 35.

Nevo R, Charuvi D, Tsabari O, Reich Z. 2012. Composition, architecture, and dynamics of photosynthetic apparatus in higher plants. The Plant Journal. 70, 157-176.

Nguyen L-T, Schmidt HA, Haeseler VA, Minh BQ. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 32, 268-274.

Novacek MJ, Wheeler QD. 1992. Extinct taxa: Accounting for 99.999...% of the earth's biota. Pages 1–16 *in* Extinction and phylogeny. Columbia Univ. Press, New York.

Nozaki H, Iseki M. 2007. Phylogeny of primary photosynthetic eukaryotes as deduced from slowly evolving nuclear genes. Mol. Bio. Evol. 24, 1592-1595.

Maddison WP, Maddison DR. Mesquite: a modular system for evolutionary analysis. 2017. Version 3.2.

Maier RM, Neckermann K, Igloi GL, Kössel H. 1995. Complete sequence of maize chloroplast genome: Gene content, hotspots of divergent and fine tunning of genetic information by transcript editing. Journal of molecular biology. 251, 614-628.

Martin W, Hermann RG, 1998. Gene transfer from organelles to the nucleus: How much, what happens and why? Plant physiology. 118, 9-17.

Martin W. *et al.* 2002. Evolutionary analysis of *Arabidopsis*, cynobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cynobacterial genes in the nucleus. Proc. Natl. Acad. Sci. USA. 99, 12246-12251.

McFadden GI, Reith ME, Munholland J, Lang-Unnasch N. 1996. Plastid in human parasites. Nature. 381, 482.

Millen RS, Olmstead RG, Adams KL, Palmer JD, Lao NT, *et al.*, 2001. Many parallel losses of *infA* from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. Plant Cell. 13,3, 645-658.

Mora C. Tittensor DP, Adl S, Simpson AGB, Worm B. 2011. How many species are there in Earth and the ocean? PLoS Biology. 9, 8.

Mukherjee AK, Ratha S, Dhar S, Debata AK, Acharya PK, Mandal S, Panda PC, Mahapatra AK. 2010. Genetic relationships among 22 taxa of bamboo revealed by ISSR and EST-based random primers. Biochem Genet. 48, 1015–1025.

Muneer PMA, Sivanandan R, Gopalakrishnan A, Basheer VS, Musammilu KK, Ponniah AG. 2011. Development and characterization of RAPD and microsatellite markers for genetic variation analysis in the critically endangered yellow catfish Horabagrus nigricollaris (Teleostei: Horabag- ridae). Biochem Genet. 49, 83–95.

Ochoa de Alda JA, Esteban R, Diago ML, Houmard J. 2014. The plastid ancestor originated among one of the major cyanobacterial lineages. Nat Commun. 15, 5, 4937.

Ohyama K. *et al*. 1986. Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. Nature 322, 572–574.

Oldenburg DJ, Bendich AJ. 2004. Most Chloroplast DNA of Maize Seedlings in Linear Molecules with Defined Ends and Branched Forms. JMB. 335, 953-970.

Palmer JD, 1983. Chloroplast DNA exists in two orientations. Nature. 301, 92-93.

Pfannschmidt T, Nilsson A, Allen JF. 1999. Photosynthetic control of chloroplast gene expression. Nature. 397, 625-628.

Pląder W. Yukawa Y, Sugiura M, Malepszy S. 2007. The complete structure of the cucumber (*Cucumis sativus* L.) chloroplast genome: its composition and comparative analysis. Cellular and molecular biology letters. 12, 584-594.

Ponce-Toledo RI, Deschamps P, Lopez-Garcia P, Zivanovic Y, Benzerara K, Moreira D. 2017. An early-branching freshwater cynobacterium at the origin of plastids. Curr. Biol. 27, 386-391.

Powell W, Morgante M, Andre C, Hanafey M, Vogel J, Tingey S, Rafalski A. 1996. The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. Molecular Breeding. 2, 225–238.

Prevost A, Wilkinson MJ. 1999. A new system of comparing PCR primers applied to ISSR fingerprinting of potato cultivars. Theoretic and Applied Genetics. 98, 107–112.

Rajkumar S, Singh SK, Nag A, Ahuja PS. 2011. Genetic structure of Indian valerian (*Valeriana jatamansi*) populations in Western Himalaya revealed by AFLP. Biochem Genet. 49, 674–681.

Rasmussen MD, Kellis M. 2007. Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. Genome Res. 17, 1932–1942.

Reith M, Munholland J. 1995. Complete nucleotide sequence pf the *Porphyra purpurea* chloroplast genome. Plant. Mol. Biol. Rep. 13, 333-335.

Rodriguez-Ezpeleta N, Brinkmann H, Burey SC, Roure B, Burger G, Loffelhardt W, Bohnert HJ, Philippe H, Lang F. 2005. Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. Curr Biol. 15, 14, 1325-1330.

Rubin GM. *et al*. 2000. Comparative genomics of eukaryotes. Science. 287, 2204-2215.

Sakamoto W, Miyagishima SY, Jarvis P. 2008. Chloroplast biogenesis: control of plastid development, protein import, division and inheritance. *Arabidopsis Book* 6, e0110.

Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature. 497, 327-331.

Salichos L, Stamatakis A, Rokas A. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. Mol. Bio. Evol. 31,5.1261-71.

Sanchez-Baracaldo P, Raven JA, Pisani D, Knoll AH. 2017. Early photosynthetic eukaryotes inhabited low salinity habitat. PNAS. 201620089.

Shinozaki K *et al.* 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO 5, 9, 2043-2049.

Soo RM, Hemp J, Parks DH, Fischer WW, Hugenholtz P. 2017. On the origin of oxygenic photosynthesis and aerobic respiration in cynobacteria. Science 255, 1436-1440.

Stevens PF. 2001. Angiosperms phylogeny websites.

Stiller JW. 2007. Plastid endosymbiosis, genome evolution and the origin of green plants. Trends Plant Sci. 12,9, 391-396.

Stiller JW, *et al.* 2003. A single origin of plastid revisited: convergent evolution in organellar genome content. J. Phycol. 39, 95-105.

Stoebe B, Kowallik K.V. 1999. Gene-cluster analysis in chloroplast genomics. Trends Genet. 15, 344-347.

Sugita M, Svab Z, Maliga P, Sugiura M. 1997. Targeted deletion of *sprA* from the tobacco plastid genome indicates that the encoded small RNA is not essential for pre-16S rRNA maturation in plastids. Molecular Genomics and Genetics. 257,1, 23-27.

Terakami S, Matsumura Y, Kurita K, *et al*. 2012. Complete sequence of the chloroplast genome from pear (*Pyrus pyrifolia*): genome structure and comparative analysis. Tree Genetics & Genomes. 8: 841

Tessier C, David J, This P, Boursiquot JM, Charrier A. 1999. Optimization of the choice of molecular markers for varietal identification in *Vitis vinifera L*. Theoretical and Applied Genetics. 98, 171–177.

Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R *et al*. 2017. GeSeq-versatile and accurate annotation of organelle genomes. Nucl Acids Res. 45 (W1), W6-W11.

Tonti-Filippini J, Nevil PG, Dixon K, Small I. 2017. What can we do with 1000 plastid genomes? The Plant Journal. 90, 808-818.

Ueda M, Kuniyoshi T, Yamamoto H, Sugimoto K, Ishizaki K, Kohchi T, Nishimura Y, Shikanai T. 2012. Composition and physiological function of the chloroplast NADH dehydrogenase-like complex in Marchantia polymorpha. Plant J. 42, 683–693.

Uyeda JC, Harmon LJ, Blank CE. 2016. A comprehensive study of cyanobacterial morphological and ecological evolutionary dynamics through deep geologic time. PLoS One. 11, e0162539.

Walker JF, *et al.* 2015. Sources of inversion variation in the small single copy (SSC) region of chloroplast genomes. American journal of botany. 102, 1751-1752.

Waller RF, *et al*. 1998. Nuclear-encoded proteins target to the plastid in *Toxoplasma gondii* and *Plasmodium falciparum.* Proc. Nat'l Acad. Science 95, 12352-12357.

Wyman SK, Jansen RK, Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics. 20, 3252-3255.

Whatley JM. 1978. A suggested cycle of plastid developmental interrelationships. New Phytol. 80, 489-502.

Wicke S, Schneeweiss GM, 2015. Next-generation organellar genomics: potentials and pitfalls of high-throughput technologies for molecular evolutionary studies and plant systematics. Next generation sequencing plant systematics. Obereifenberg; Koeltz, Botanical Book.

Wicke S, *et al.* 2011. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol. Biol. 76, 273-297.

Wolfe KH, Morden CV, Palmer JD, 1992, Function and evolution of a minimal plastid genome from a nonphotosynthesis plant. Proc. Nat'l Acad. Science 89, 10648-10652.

Yoon HS, *et al.* 2004. A molecular timeline for the origin of photosynthetic eukaryotes. Mol. Bio. Evol. 21, 809-818.

Zhang K, Zhu X, Wood RA, Shi Y, Gao Z, Poulton SW. 2018. Oxygenation of Mezoproterozoic ocean and the evolution of complex eukaryotes. Nat Geosci.

Zhang N, Zeng L, Shan H, Ma H. 2012. Highly conserved lowcopy nuclear genes as effective markers for phylogenetic analyses in angiosperms. New Phytol. 195, 923–937.